JESLA Journal of the European Second Language Association

Lindstromberg, S. and Eyckmans, J. (2017). The particular need for replication in the quantitative study of SLA: A case study of the mnemonic effect of assonance in collocations. *Journal of the European Second Language Association,* 1(1), 126–136, DOI: https://doi.org/10.22599/jesla.26

SYSTEMATIC REVIEW AND META-ANALYSES

The particular need for replication in the quantitative study of SLA: A case study of the mnemonic effect of assonance in collocations

Seth Lindstromberg* and June Eyckmans†

Recent surveys of published reports of quasi-experimental studies of second language acquisition (SLA) indicate that low statistical power is pervasive owing in large part to small average sample sizes. The surveys do not indicate a marked trend toward samples that are larger. After illustrating the problem of low power in SLA research, we review arguments that increased replication of original studies can enable small-sample quantitative researchers to make firmer contributions to the field of SLA, especially if estimation of effect sizes and the practice of on-going statistical meta-analysis become routine. As a case study, we describe a series of small-sample quasi-experiments of which the first five found a short term positive mnemonic effect of interword, intra-phrase vowel repetition (or assonance) on learners' retention of the forms of L2 collocations (e.g. strong bond vs. firm hold), whereas a sixth study newly reported here found negative effects. The case study illustrates the roles of replication and meta-analysis in successive re-adjustments of an original estimate. More specifically, the case study illustrates a meta-analytic approach to making sense of conflicting outcomes. All in all, it illustrates why small-sample researchers need to adopt a more long-term view.

Keywords: Small-scale quantitative research; Meta-analysis; Replication; L2 phrase retention

Publisher's Note

In a latter stage of manuscript preparation the key term '(quasi)experimental study' was changed to 'quasi-experimental study', and the same for the plural. The former term, which is meant to include quasi-experiments and true experiments, indicates the scope of the article whereas the second term does not.

1. Introduction

It has been pointed out that in the field of SLA, replications of quasi-experimental and other quantitative studies rarely find their way into print, whereas in more mature fields (e.g. the natural sciences) an experimental finding is likely to be given little or no weight until it has been replicated and the results made known (Mackey, 2013; Porte, 2012a). With so much recent attention given to the issue of sample size (the number of learners per treatment group), we consider the need for more replication studies and discuss a profitable use of the results. Throughout this article, we focus exclusively on 'external' replication, whereby an original study is in key respects re-conducted by the same or different researchers so that one or more of the original research questions may be re-addressed with recourse to inferential statistics, but also with data being collected from new study participants or, conceivably, from the same participants but with respect to new items. (For variations of this definition and for classifications of replication studies see articles in Porte 2012b, especially Polio 2012a). Although our examples and associated discussion relate mainly to quasi-experimental investigations of pedagogical interventions in the context of instructed SLA, most of what is said applies equally to small-sample observational studies. We begin by arguing that replication studies are especially necessary in the quantitative study of SLA because of small average sample sizes. We then present a case study illustrating the practice of small-scale 'Continuously Cumulating Meta-Analysis' (Braver, Thoemmes, & Rosenthal, 2014; cf., Asendorpf et al., 2013; Cumming, 2012, 2014), an approach to synthesizing the results of multiple related studies that affords small-sample researchers with enhanced possibilities of drawing credible and precise conclusions from research results. In what follows we often refer to 'small samples'. What we mean are studies in which the number of observations (e.g. test scores) per sample is less than 30 or so. In contrast, a large-sample study would be one in which the number of observations

* Hilderstone College (emeritus), UK

† Ghent University, Department of Translation, Interpreting and Communication, BE

Corresponding author: Seth Lindstromberg (sethl@hilderstone.net)

were several times that number.[1] We propose these benchmarks, despite their conspicuous elasticity, solely for the purpose of orienting our discussion in a general way; a judgment of study size for more specific purposes must take account of a wide range of factors besides sample size (e.g. the reliability of measurements, experimental design, and the method of statistical analysis).

2. Statistical Power, Sampling Variation, and Meta-analysis in Quantitative Studies of SLA

2.1. Typical sample sizes in quantitative studies in SLA

Focusing on observational and quasi-experimental studies relating to Long's (1983) interaction hypothesis, Plonsky and Gass (2011) surveyed 174 articles in 14 applied linguistics journals (plus two edited volumes) from the period 1980 to 2010, finding that the mean number of learners per learning condition was 22. In a survey embracing 606 quantitative studies published in two top applied linguistics journals, Plonsky (2013) found that the median number of learners per learning condition was 19. In a survey of 96 reports of quasi-experimental studies of instructed SLA that appeared in one journal from 1997 to 2015, Lindstromberg (2016) found that the median per condition number of learners in within-subjects designs was 26 but across the much more common studies with between-subjects and mixed designs the median sample size was 20. Lindstromberg (2016) found, moreover, that the median sample of stimulus items was just 15. We now turn to the matter of statistical power and how it is affected by a reliance on small samples of learners.

2.2. Implications of small sample sizes for statistical power

When we say that a study has low (statistical) power, we mean that the average power of its tests of statistical significance is low assuming (as we do throughout this article) that $\alpha = 0.05$ and all significance tests are two-tailed. With respect to any particular significance test that is to be used on a given set of data, having low power means that there is an undesirably high probability that the test will find $p > \alpha$ even when a hypothesized effect actually exists (e.g. Ellis, 2010). Prominent among the factors which determine power are the sizes of samples and the sizes of true effects, that is, the effects in the relevant population(s) from which samples are drawn. Importantly, true effects are influenced by population variances (or SDs).

As it happens, low power has been identified as a serious problem in SLA quantitative research (Plonsky, 2013). In essence, true effects are nearly as likely to be overlooked as to be detected. To explore the issues here, we carried out a number of computer simulations of the statistical power deployed in situations typical for our field when n is set at 20, Student's independent samples (IS) t-test is used, and parameters are set so that all assumptions of the IS t-test are satisfied (e.g. the parent populations were normal with equal variances and the tested samples were of the same size). **Figure 1** shows histograms summarizing the results of two of the simulations.[2] The histogram on the left illustrates the power of the IS t-test given the conditions just described when $d = 0.50$. As can be seen, even though the SDs are likely to be untypically small, the IS t-test affords only a 34% chance to detect a true effect that is large enough to have solid practical significance in many circumstances (e.g. Ellis, 2010; Grissom & Kim,

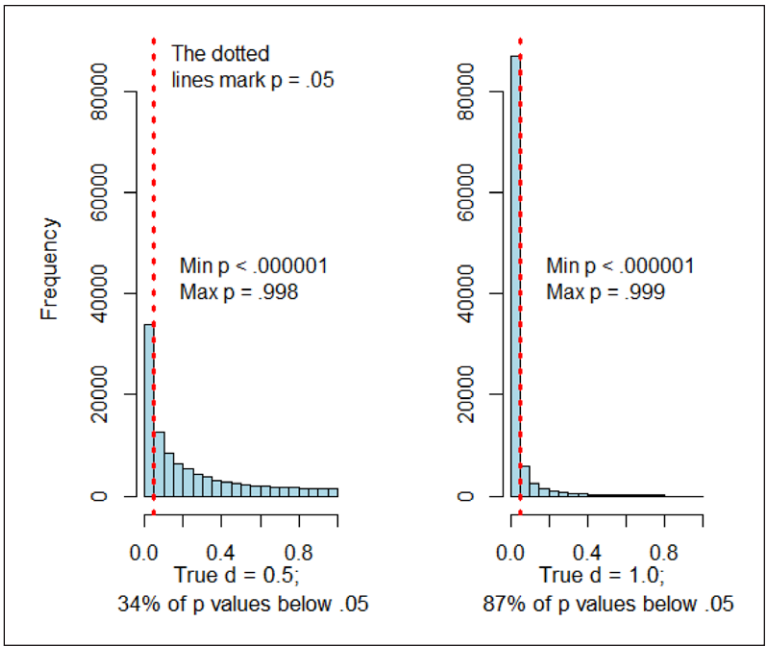

Figure 1: Histograms showing the distribution of p values from 100,000 t-tests of the difference between a simulated sample X and a simulated sample Y. For the histogram on the left, 100,000 X samples and 100,000 Y samples were randomly drawn from normally distributed populations (both $SD = 1$) where $Mn._{\text{Population.X}} = 10$ and $Mn._{\text{Population.Y}} = 10.5$. For the plot on the right $Mn._{\text{Population.Y}}$ was changed to 11.0.

2012, pp. 127–130).[3] The histogram on the right of **Figure 1** illustrates one mathematically coherent solution to the power problem confronting the small-sample SLA researcher in that it shows how dramatically power can be improved when the true effect is 1.00 rather than 0.50. However, it is unlikely to be reasonable or even possible for researchers to investigate only effects as large as this. For one thing, large effects are by no means the only effects that are practically and/or theoretically important. For another, in SLA research it does not tend to be easy to estimate true effect sizes in advance of conducting a study, a reason for this being that serviceable data such as previously observed effect sizes can be difficult to find in the literature owing to patchy reporting (e.g. Plonsky, 2013). To illustrate a second mathematically coherent remedy, let us refer again to the simulation summarized in the leftmost histogram in **Figure 1**. If this simulation were re-run with *n* raised from 20 to 64, power would rise to 0.80, which is often said to be the minimum acceptable level in many situations (Ellis, 2010). However, surveys of SLA research reports have noted no conspicuous change in the average sample size in recent decades (e.g. Lindstromberg, 2016); and it is not clear what realistic development could cause certain well-known practical constraints on sample sizes (such as the typical size of language classes) to relax so much that average sample sizes in SLA research can any time soon become two or three times as big as they are now.

Still, there are various other actions that researchers can peform in order to deploy more statistical power besides increasing sample sizes (or raising α). One such action is to make appropriate use of modern, robust statistical procedures such as bootstrapping (Wilcox, 2005). Another is to make increased use of optimum experimental designs (McClelland, 1997; Westfall, Kenny, & Judd, 2014). A third is to adopt Bayesian methods (Dienes, 2014). There is no doubt that these three options have been underused by SLA researchers: for example, Lindstromberg (2016) found no mention of any of them in the 96 articles he surveyed. However, it is not clear how much potential an increase in use of robust methods and optimum experimental designs has to bring about a huge improvement in field-wide average statistical power from the current low level of about 0.57 (Plonsky, 2013) to 0.80 or, preferably, higher. And much the same thing might be said of other mainstream practices with the potential to raise power that may be discussed in textbooks (e.g. Baguley, 2012). (We return to Bayesian methods near the end of this article.)

2.3. What an initial p value foretells about a replication p value

Returning to **Figure 1**, each histogram displays an empirical sampling distribution of *p* over many ideally exact replications in which everything stays the same except the participants. It can be seen in this figure that the range of *p* in both distributions is virtually equal to its mathematical maximum, 0—1, regardless of whether $d._{\text{Population}}$ is 0.50 or 1.00. Cumming (2008, 2012) has extensively discussed the sampling variation of *p* in relation to the issue of replication. Cumming (2008) showed that a researcher who observes $p = 0.05$ in an initial experiment has no reason to be confident that a replication will find a significant *p* value, whatever the size of *n*. He has summed up this state of affairs as follows: "A *p* value is typically a very poor measure of the strength of evidence against a null hypothesis" (Cumming, 2012, p. 134) and "anything other than a very small *p* value gives virtually no useful information at all" (p. 135).

Although researchers in general are likely to be well aware that indicators such as sample means tend to show wide sampling variation when samples are small, it has repeatedly been found that researchers of all degrees of experience, including statisticians, are likely to have such limited awareness of the fact that *p* values show similarly wide variation that they greatly overestimate the probability that a significant *p* value from an original study will be followed by a significant *p* value in a procedurally identical replication (e.g. Lai et al., 2012; Tversky & Kahneman, 1971). Lai et al. note that certain optimistic statements about the replicability of an initial *p* value that have been put forward in the technical literature of applied statistics are based on two generally unwarranted assumptions: first, that samples will be large and, second, that researchers will be able to use precise estimates of population means. We have seen that the first assumption is not apt for SLA research while the second assumption, according to Lai et al. (2012), is dubious because it depends on an unlikely combination of good methods of measurement, thorough understanding of the research territory, large samples, normally distributed data, regularity of variances (i.e. homoscedasticity), and good luck.

The point of the remarks directly above is that quantitative researchers of SLA may well be as likely as researchers in other fields to overestimate the probability that an original study's significant *p* value will be followed by a significant *p* value in a replication. A danger here is that a tendency to put too much faith in significant *p* values may reduce researchers' motivation to conduct replications. It would therefore be highly convenient if there were measures that are better indicators than *p* for whether a given experimental finding is trustworthy. Unfortunately, possible alternatives are subject to sampling variation as well. To take *d* for instance, not only can values of *d* show wide sampling variation when *n* is small but (unlike *p*) *d* can fall on either side of zero. This means, for example, that when there is a true positive effect, an original study may find a positive value of *d* whereas a replication might find a negative value (or vice versa). All in all, a single small-sample study such as ones that abound in quantitative SLA research cannot provide a precise estimate of an effect's size or, quite possibly, even its direction. Naturally, a large-scale random controlled trial (RCT) is superior in this regard. However, findings of recent surveys of the SLA research literature (e.g. Plonsky & Gass, 2011; Plonsky, 2013; Lindstromberg, 2016) do not suggest that the community of SLA researchers has the resources to conduct enough large-scale RCTs to address a good-sized proportion, let alone all, of the many worthwhile hypotheses that have been, and are being, generated. Finally, we have looked at the sampling

variation of *p* values derived from the IS *t*-test. Since ANOVA is the other workhorse of quasi-experimental studies in SLA (Gass, 2009; Lindstromberg, 2016), it should be noted that *p* values deriving from ANOVA are especially unstable from study to study (Grissom & Kim, 2012, p. 183).

2.4. Effect sizes in quantitative studies of SLA

For some decades now quantitative researchers have been urged to calculate, report, and verbally interpret observed effect sizes (e.g. Cohen, 1994; Ellis, 2010). Journal surveys (e.g. Lindstromberg, 2016) indicate that when SLA researchers do report and interpret effect sizes they generally fall back on Cohen's well-known default benchmark interpretations. For *d* these interpretations are: 0.20 indicates a small effect that may nevertheless have substantive significance; 0.50 indicates a medium effect; and 0.80 indicates a large effect. However, Plonsky and Oswald (2014) found that that effect sizes reported in quantitative studies of SLA have tended to be larger than Cohen's benchmarks would lead one to expect. Going by the median and the inter-quartile range of the effect sizes they observed in the articles they surveyed, Plonsky and Oswald (2014) proposed new, field-specific benchmarks. For between-subjects studies these are: 0.40 = small, 0.70 = medium, and 1.00 = large. These researchers acknowledged, however, that their new benchmarks were based on a distribution of reported effect sizes likely to have been inflated by the well-documented publication bias in favor of reports of statistically significant results. Thus, it is not necessarily true that Cohen's benchmarks are inappropriate for research in SLA. In any case, SLA researchers are likely to be aware that a given numerical measure of effect size can refer to effects of differing substantive significance depending on the context in which each effect is manifest. It may be worth adding, though, that when a phenomenon occurs repeatedly or when a process is permanently on-going in such a way that its effect accumulates (which is likely to be a common state of affairs in language use and language learning), an estimate of effect size that is derived by use of a conventional formula may portray the effect as very small (e.g. $\omega^2 = 0.003$) even though it is of clear, perhaps even major, substantive importance (Abelson, 1985; see also Breaugh, 2003).

2.5. Statistical meta-analysis: a way forward for small-sample quantitative research

The wide variation in raw and standardized estimates of effect size that is likely to be seen across studies addressing the same or similar research questions (e.g. Cumming, 2012) could be discouraging were it not for the fact that a series of such estimates can be averaged, by means of statistical meta-analysis, to yield a 'pooled' estimate that is likely to be more precise – hence, bracketed by a narrower confidence interval – than an estimate stemming from any individual study. Moreover, because a pooled estimate derives from a widened database, it is generally more credible than an estimate based on any one of the individual studies at issue. An exception to this generalization would be a high *N*, high quality RCT – at least when the alternative is a comparatively low *N* meta-analysis based on a handful of individual studies. But we have mentioned that the option of conducting a high *N* RCT is rather unlikely to be a practical option in our field. In any case, wondering whether a meta-analysis is better than a RCT is unfruitful since neither option excludes the other. For instance, RCTs can be included in meta-analyses. An additional advantage of a meta-analysis is that it mathematically utilizes previous estimates of effect size, whereas a RCT does not. Finally, there are situations in which it may be worthwhile to conduct a meta-analysis on just two studies – provided they are sufficiently similar and if no additional suitable studies are available – since even a meta-analysis this small can yield an estimate of effect size that is considerably more accurate, precise, credible, and useful than the individual estimates from the two studies (Cumming, 2012, pp. 181–186; Valentine et al., 2010). That said, the more studies the better (e.g. Borenstein et al., 2009; Valentine et al., 2010). In short, meta-analysis on any scale can enhance the contributions of small-sample quantitative research to the development of shared knowledge—provided, of course, that the studies that get meta-analyzed are reasonably well-conceived, well-designed, and well-conducted, and that they either report observed effect sizes or else provide the statistics such as means and SDs that enable meta-analysts to calculate effect sizes retrospectively.

3. Case Study: Does Assonance Make the Forms of L2 Multiword Units Relatively Easy to Recall?

3.1. The basis of the hypothesis that assonance can have a practically significant mnemonic effect

It is now known that learners wishing to attain a high level of proficiency in an additional language must learn not just thousands of words but also thousands of recurrent word phrases, or multiword units (MWUs), including situational formulae such as *fingers crossed*, figurative idioms such as *cut corners*, discourse markers such as *by the way*, and strong collocations such as *commit a crime* (see articles in Polio, 2012b). It is also known that post-childhood learners find this difficult to achieve (e.g. Forsberg, 2010; Laufer & Waldman, 2011; Nekrasova, 2009). A number of proposals have been made about how learners can be helped to accelerate the rate at which they acquire these phrases (for a review, see Boers and Lindstromberg, 2012). With regard mainly to English, one of these proposals is based on the observation that a substantial proportion of English MWUs—especially ones whose meaning is or can be figurative—show either alliteration (e.g. *face the facts*) or assonance, that is, intra-phrase, inter-word vowel repetition in prominent syllables of content words, as in *blow your nose*. Relevantly, there is very strong evidence that alliteration occurs above baseline, or chance, rates in figurative MWUs of a variety of types (Boers and Lindstromberg, 2009; Gries, 2012; Lindstromberg, Forthcoming) and strong evidence that the same is true of assonance (Lindstromberg, Forthcoming). Assuming, then, that the stock of English MWUs manifests a surplus of assonance, one may ask whether the extra assonance in these expressions has a function. One

hypothesis (Lindstromberg, Forthcoming; cf. Gries, 2011) is that assonance, along with rhyme and alliteration, can be sufficiently perceptually or cognitively salient that it can serve to draw attention to MWUs that are particularly likely to be used at key junctures in discourse. Consistent with this hypothesis is McCarthy's (1998) finding that speakers and writers are particularly likely to use figurative idioms to deliver summative evaluations of an event or state of affairs. In light of the above, and given the present authors' ongoing interest in the instructed acquisition of L2 MWUs, the research questions that we have investigated concern whether alliteration and assonance make L2 English MWUs easier to recall and, if so, under what circumstances—our ulterior goal being that of finding out whether there are steps that teachers can take to help learners acquire alliterative or assonance MWUs more quickly than would otherwise be the case. As already indicated, in this article we focus on assonance. For reviews of studies having to with alliteration see Boers and Lindstromberg (2012) and Boers, Lindstromberg and Eyckmans (2014).

3.2. Previous studies

A number of quasi-experimental studies have been carried out to determine, firstly, whether assonance actually does make the forms of MWUs relatively easy to remember and, if so, how much easier and under what circumstances (Lindstromberg & Boers, 2008; Lindstromberg & Eyckmans, 2014).[4] An issue that was addressed in some studies was the effect that raising awareness, in combination with an attention direction task, may have on retention of form. The studies we focus on below involved MWUs deemed highly likely to be familiar to and understood by the mostly advanced proficiency participants. A rationale for using familiar stimulus expressions is that doing so may simplify the task of identifying effects associated with phonological form by largely eliminating effects arising from semantic or phonological/orthographic novelty.[5] In each of the studies we refer to, non-native speaking adults were asked to study a set of English phrases before being tested by, for example, hearing them, repeating them chorally, writing them down, and then sorting them. The great majority of these targeted items are adjective-noun and noun-noun phrases which consultation of the Corpus of Contemporary American English showed to be among the most frequent collocations of at least one of the two constituent words. In each study, half of the stimulus collocations assonate (e.g. *strong bond*) and half are ones that neither assonate nor alliterate (e.g. *firm hold*). Steps were always taken to ensure that the collocations in each of the two sets are similar in terms of variables such as frequency, number of syllables, and concreteness-imageability of meaning. Shortly after the study phase of each experiment, the participants were tested on their ability to recall the forms of the collocations they had studied. They were usually tested again after a delay of either a day or a week. The (near) immediate post-tests that followed the study phase involved either free or cued recall. In the latter case the cue was always the initial noun or adjective word of a target collocation. The delayed post-tests usually involved only cued rather than free recall. There were, additionally, two delayed post-tests of recognition. The key research questions were:

> RQ1: Are the forms of assonant collocations relatively well remembered?
> RQ2: If so, to what degree?
> RQ3: Is the type of task performed in the study phase associated with a difference between the retrievability of the assonant and control collocations?
> RQ4: If so, to what degree?
> RQ5: How durable is any extra retrievability of the assonant collocations?

A positive mnemonic effect of assonance, even one that endures for only a few minutes, could have practical significance if the following scenario is realistic, which we believe it is. That is, a learner hears a L2 MWU. Later, perhaps in an ongoing conversation, the learner not only remembers the earlier encounter with the MWU but also remembers something of the context in which it was used (e.g. who the speaker and addressee were and what the topic was). By virtue of remembering the earlier encounter, the learner is better equipped to interpret the MWU on any subsequent encounter (if the meaning of the MWU was previously unknown) and may also be better equipped than before to use it autonomously, especially if previously the form of the MWU was shallowly entrenched in the learner's memory. Finally, the presence of assonance may raise the chances that the scenario just outlined takes place to the extent that it confers perceptual or cognitive salience. (For additional rationale see Boers and Lindstromberg, 2009, 2012.)

3.3. Meta-analysis

Four of the experiments discussed above (Lindstromberg & Boers, 2008; Boers et al., 2014, experiment 2; Lindstromberg & Eyckmans, 2014, experiments 1 and 2) were included in a small meta-analysis reported by Lindstromberg and Eyckmans (2014). Those four experiments targeted a total of 51 different collocations (some of which were used in more than one experiment) and involved 121 different post-childhood learners of English as a foreign language. As mentioned, the purpose of the meta-analysis was to obtain an improved estimate of the mnemonic effect of assonance on short term recall of the forms of two-word collocations *after a study phase involving some kind of deliberate direction of attention to phonological form.* Accordingly, the meta-analysis only considered results from near-immediate post-tests of free or cued recall—that is, tests administered within ten minutes of the conclusion of the treatment phase.[6] Thus, one previously reported assonance experiment (Boers et al., 2014, experiment 1) was not included because its study phase involved no direction of attention to phonological form. The pooled estimate yielded by the meta-analysis was $d = 0.51$. Results from the various *delayed* post-tests of the experiments just referred to were not subjected to meta-analysis on account of their incomparability. (One of those tests probed recognition; of the other two tests,

one was given after a delay of a day whereas the other was given after a week.)

3.4. A new experiment

Each of the experiments mentioned above was intended to address the following two questions at least: Do assonant collocations tend to be better remembered than similar control collocations? If so, how much better? Because all of these experiments were small in scale, we must assume that their results were particularly susceptible to the vagaries of sampling variation. Moreover, because they were real-world replications rather than computer simulations, there were bound to be additional sources of variation such as random experimental error. Accordingly, it seemed prudent to gain an additional result regarding the mnemonic effect of assonance after a treatment featuring overt attention direction.

The 81 participants were enrolled in their first year of a bachelor's programme in applied linguistics at the University College of Ghent (Belgium). Their mother tongue was Dutch, and English was one of two foreign languages in their programme. Their proficiency in English was estimated to be level B2 according to the descriptors of the Common European Framework of Reference, which corresponds to an IELTS score of at least 5. Twenty-three students were not present for the delayed post-tests. All participants supplied written consent.

A set of 28 stimulus collocations was prepared, 14 of which show assonance and 14 of which (the 'controls') show neither assonance nor alliteration. These collocations were drawn from the set of 32 used by Lindstromberg and Eyckmans (2014). To increase inter-collocation similarity, four collocations which raters had judged to be especially abstract in meaning were omitted from the new set of 28. As can be seen in **Table 1**, the two sets were closely balanced in terms of two variables with potential to affect memorability – frequency and concreteness-imageability (CI). Not shown in the table is the balance between assonant and control expressions with respect to the frequency of the rightward collocates (mean frequencies: $62{,}592_{assonant}$ vs. $56{,}833_{control}$). In tests of recall from lists of words that vary widely in frequency, the correlation between recall and frequency is generally *negative*, although not necessarily strongly so. Also not shown in **Table 1** are the mutual information scores, that is, the indices of the statistical association between the left and right collocates (mean scores: 5.7_{asson} vs $4.9_{control}$). The influence of this type of statistical association on the recall of L2 collocations does not appear to be known, although there is some evidence that it is not strong (see Lindstromberg & Eyckmans, 2014). Finally, the assonant and control expressions were approximately balanced in terms of a number of formal variables (e.g. length in terms of syllables and phonemes), additional semantic variables such as emotiveness and personal relevance, and syntactic structure.

The treatment began with a dictation of the 28 collocations. Subsequently, the concept of assonance was explained to the participants in terms of vowel repetition in content words. Then they were given a randomized list of the 28 test items and asked to circle ones not showing assonance. The rationale for this task was that in the experiment of Lindstromberg and Boers (2008), where a strong positive effect of assonance had been observed, participants' attention had been directed to occurrences of assonance. In case those participants' superior recall of the assonant collocations had come about because of a task effect rather than because of an effect of assonance, in all subsequent experiments with attention direction tasks the instructions for these tasks focused on the control collocations.

Following the treatment, in order to disrupt possible attempts to rehearse the studied collocations, participants were asked to change seats. The instructor then tested free recall by asking participants to write down as many of the collocations as they could remember. A week later, an unannounced test of free recall was administered to the 58 students who had attended the treatment sessions. Descriptive statistics for by-participants scores are given in **Table 2**. Against expectation, the non-assonant collocations were better recalled in both of the posttests even though the opposite result had been obtained in

Table 1: The 28 stimulus collocations used in the new experiment.

The assonant collocations[a] (n = 14)	The control collocations[a] (n = 14)
town house[1], *gift list*[2], *deep sea*[3], *quick trip*[4], *land mass*[5], *main gate*[6], *soft cloth*[7], *loud sound*[8], *safe place*[9], *job loss*[10], *high price*[11]	*town square*[1], *check list*[2], *deep hole*[3], *quick stop*[4], *land use*[5], *main road*[6], *soft ground*[7], *sharp sound*[8], *nice place*[9], *heat loss*[10], *high rate*[11]
strong bond[A] , *gas tank*[B] *rubber glove*[C]	*firm hold*[A], *tool box*[B], *metal roof*[C]

Mean (% greater)	Variables	Mean (% greater)
2.33 (7.4%)	Concreteness- imageability rating	2.17
305	Whole collocation frequency	306 (0.3%)
125,978	Frequency of both words combined	137,931 (9.5%)
5.7 (16.3%)	Mutual information score	4.9

Note. The superscript Arabic numbers (e.g. in *high price*[11], *high rate*[11]) indicate assonant and control collocations that share a word. The superscript capital letters (e.g. *strong bond*[A], *firm hold*[A]) indicate matchable collocations that do not overlap in this way but which do show syntactic and semantic similarity.

two previous experiments featuring the same attention-direction task (Lindstromberg & Eyckmans, 2014).

We turn now to the inferential analysis. The assonance studies referred to so far (Boers, et al., 2014; Lindstromberg & Boers, 2008; Lindstromberg & Eyckmans, 2014) employed a traditional analysis of scores organized by participants. A disadvantage of an analysis based only on by-participants scores is that such an analysis fails to take full account of variation in scores across targeted vocabulary items (by-items scores). In inferential analysis of data from studies with our experimental design, application of mixed-effects modelling (a relatively new approach in our field) enables one to take better account of the information in the data. Consequently, a mixed-effects approach can provide a basis for wider and sounder generalizations than can legitimately be made on the basis only of by-participants (or by-items) scores (e.g. Baguley, 2012, Linck & Cunnings, 2015).

To implement the mixed-effects analysis, we used the open source statistical computing environment *R* (R Core Team, 2016) and the glmer function in the R package *lme4* (version 1.1–12) to test generalized linear mixed-effects logistic regression models fitted by maximum likelihood with Laplace approximation (Bates, Mächler, Bolker, & Walker, 2017). In these models, the dependent variable was Recall as measured by the test scores. The independent variable, Type of Collocation (+/– assonance), was treated as a fixed effect. Intercepts for individual learners and for individual collocations were allowed to vary by Type of Collocation. **Table 3** summarizes the results, from which it can be seen that the superior recollection of the control collocations is statistically significant in neither post-test. (A slightly more complex mixed-effects model in which slopes were also allowed to vary across learners turned out to be slightly less efficient than the model we report.) Because the regression model we used is a type of logistic regression, the coefficients given in column 2 of **Table 3** are expressed in terms of logits (or logged odds). By exponentiating, or taking the anti-log of, the coefficient for a variable we obtain an odds ratio (OR) (column 4 of **Table 3**), which serves as our measure of effect size. The OR of 0.91 (for the near-immediate post-test) means that on average the odds of a learner recalling any given assonant collocation are 91% as high as that learner's odds of recalling any given non-assononant collocation. According to a formula given by Grissom and Kim (2012, p. 273), this is roughly equivalent to $d = –0.05$. For the delayed post-test, OR = 0.79 ≈ $d = –0.13$. Effects of these sizes are generally considered to be very small.

If we were now to focus solely on the *p* values associated with the negative results shown in **Table 3**, we could not say much more than that these results are awkward for the hypothesis being tested. Adoption of the meta-analytical perspective opens up an additional, fruitful course of action. Specifically, when there appears to be no vitiating errors in experimental design or procedure (as in the present case, we believe), each new result should be regarded as evidence that can be added into an updated meta-analysis for a more credible estimate of effect. Indeed, in a meta-analytic approach to small-sample research, a complete absence of null or negative results is grounds for suspecting that one has not yet looked at a truly representative range of studies, especially when the effect at issue is likely to be of small or medium size (cf., Borenstein, et al., 2009, pp. 283–284; Ellis, 2010, pp. 120–122). In general then, meta-analysis is a continuing process whereby successive estimates of effect size show greater precision (Braver et al., 2014; Cumming, 2012). Let us now turn to a second, new small-scale meta-analysis that included all but one of the data sets figuring in the meta-analysis reported by Lindstromberg and Eyckmans (2014) plus the immediate post-test data from the experiment newly reported here.

3.5. A new meta-analysis

Because the small-scale meta-analysis reported by Lindstromberg and Eyckmans (2014) was based on effect sizes calculated only from by-participants scores, we retrospectively applied mixed-effects modelling procedures to the immediate post-test data from the experiments at issue. This yielded regression coefficients expressed in logits. (For software, we used the R package metafor developed by Viechtbauer, 2010.) Using the coefficients from the old studies and from the study newly reported here, we were able to carry out a new small-scale meta-analysis on a firmer basis than before. Unfortunately,

Table 2: The new experiment: Immediate and delayed tests of participants' ability to recall assonant and non-assonant collocations.

	Free recall of whole collocations	
Delay after the study phase:	2–3 minutes	one week
Number of participants:	81	58
Total per-student recalls, assonant vs. control:	384 v. 409	100 v. 133
Raw *MD*s:	–.31	–.57
*SD*s of raw assonant and control scores:	2.29, 2.50	1.35, 1.48

Table 3: Key results from the mixed-effects logistic regression models for the new study.

		Coefficient [95% CI]	SE	Odds Ratio [95% CI]	*z*	*p*
Posttest	*Near immediate*	–0.10 [–0.57, 0.38]	0.23	0.91 [0.56, 1.46]	–0.42	.67
	Delayed	–0.24 [–1.01, 0.55]	0.38	0.79 [0.36, 1.73]	–0.64	.52

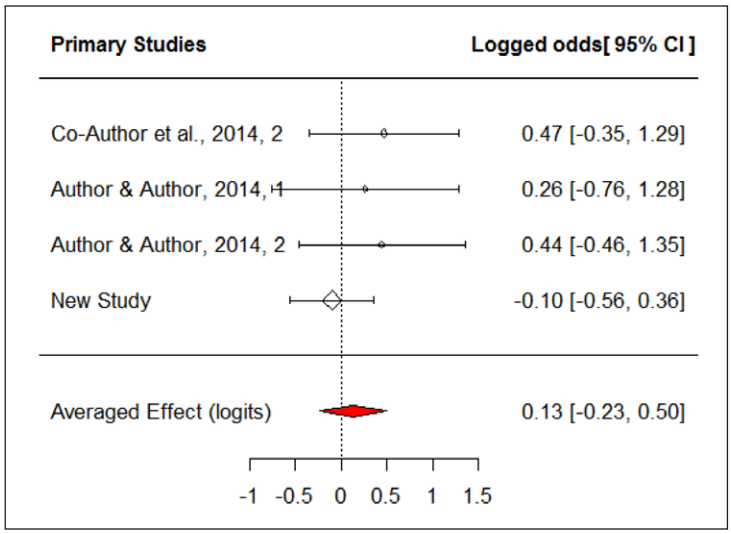

Figure 2: A forest plot summarizing a random-effects meta-analysis based on coefficients (measuring logged odds) from mixed-effects logistic regression analyses of (near) immediate posttests of recall from four primary studies. The solid diamond at the bottom shows the $CI_{95\%}$ for the pooled estimate of effect size.

the only test scores from the study of Lindstromberg and Boers (2008) to survive various subsequent changes of hardware and software were in by-participants format. Consequently, the positive effect observed in that study could not be included in the new meta-analysis.

The results of the new meta-analysis are summarized in **Figure 2**. In this figure the $CI_{95\%}$ for the pooled estimate of effect size (logit = 0.13) is indicated by lateral extent of the solid diamond near the bottom.[7] Exponentiation of this estimate and of the two limits of its CI gives OR = 1.14, $CI_{95\%}$ [.80, 1.65]. By this estimate, on average, the odds that a learner will recall a given two-word assonant collocation are 14% higher than the odds that the learner will recall a given two-word non-assonant collocation. This equates to d = 0.075, which would be considered very small in most circumstances. The wide CI for the OR straddles = 1, meaning that this meta-analysis fails to establish the direction of the true effect of assonance. Because this estimate could not take into account the positive effect found in the study by Lindstromberg and Boers (2008), it may well be somewhat low. Also, the positive point estimate (which is represented by the fattest part of the solid diamond) is about seven times more probable than the values at the ends of the CI (Cumming, 2012). Nevertheless, the estimate of the true effect of assonance given by Lindstromberg and Eyckmans (2014) ($d \approx 0.50$) now seems to be much too high.

To sum up, the available evidence leaves open the possibility that assonance has a modest short-term positive effect on the retrievability of L2 English collocations following a taskwhich directs learners' attention to phonological form.

4. Summary and Conclusion

We have presented a case study of the variation in observed effect sizes across a series of largely small-sample, quasi-experimental studies intended to estimate the direction and size of the effect of assonance on the near immediate recall of the forms of L2 collocations so that any practical importance of this effect might become evident. Initial results accorded well with the hypothesis that assonance has a positive effect on recall but eventually an experiment produced negative results. The gist of an extended meta-analysis including these results is that there may well be a positive mnemonic effect of assonance but that it is premature to suppose that it is large enough to be exploited by teachers and materials writers with the goal of facilitating the acquisition of some types of English MWUs.

Additionally, we have shown how negative results are dealt with in a meta-analytic approach to research. In this approach, especially when samples are small, it is accepted that widely dissimilar results may occur and that as a matter of course, replications are required in order for a credible and usefully precise estimate eventually to be obtained through meta-analysis.

A question of relevance throughout this article has been how to think about cases where $p > 0.05$ even though descriptive statistics (e.g mean differences) means indicate that there might well be an effect after all. Our recommended solution to the problem—namely, increased use of meta-analysis to raise statistical power high enough to obtain a sufficiently precise estimate at $p < \alpha$—is based on the mundane fact that power can be increased by increasing *N*. Indeed, nothing in the approach we have recommended conflicts with the standard practice of null hypothesis significance testing (NHST). It seems relevant to note that if NHST were abandoned, researchers would be able to focus on the sizes of effects undistracted either by concerns about statistical (in)significance, by the associated issue of statistical power, or by the possibility of finding $p > \alpha$ when an effect size $\neq 0$. As readers will be aware, abandoning NHST is a development that many authorities have called for (e.g. Cohen, 1994; Cumming, 2012, 2014; but for counter-arguments see Cortina and Landis, 2011; and see Norris, 2015 for discussion of NHST in SLA research generally). However, as a reviewer pointed

out to us, a Bayesian approach affords quite different solutions to the problem of finding $p > \alpha$ even though the effect size ≠ 0. For insightful discussion of options and opportunities in this regard, see Dienes (2014). Additionally, replication has been a constant theme in this article but even so there is vastly more than could be said about it. For relevant discussion, including many suggestions for action, see Brandt et al. (2014) and Nosek et al. (2015).

To conclude, our main point in this article has been that without replication *and* meta-analysis, the usefulness of small-sample quantitative experimental research can legitimately be called into question, but that if both of these practices become routine, small-sample researchers can make solid contributions to the field. However, in small-sample research arrival at a usefully precise estimate of a hypothesized effect is likely to demand numerous trials and multiple years of effort, with reversals of direction being all too likely along the way.

Notes

[1] For some applied statisticians the threshold of 'large' can be 200 or even more, depending on the type of test (e.g. Field, Miles, & Field, 2012, p. 175; Wilcox 2005, p. 421). Naturally, the size of one's samples of items is an important variable as well (Westfall, Kenny, & Judd, 2014).

[2] In the human sciences, generally such a favorable constellation of perfectly satisfied assumptions is comparatively unlikely, meaning that statistical power in real research is quite often lower than canonical power calculations will suggest (Micceri, 1989; Wilcox, 2017).

[3] The plots in **Figure 1** also illustrate why it tends to be unwise to cite a high *p* value as evidence that a particular effect is absent or substantively insignificant.

[4] All these tests are likely to have probed retention from *episodic* rather than *semantic* memory (Tulving, 1983). For further discussion of the relevance of this matter to the case at hand, see Lindstromberg and Eyckmans (2014). For a review of research into phonological similarity effects in experimental psychology, see Eyckmans and Lindstromberg (2016).

[5] For an account of evidence that patterns of inter-word intra-phrase phonological similarity facilitate the recollection of the forms of previously *unfamiliar* MWUs, see Eyckmans and Lindstromberg (2016).

[6] It has been argued that in order to probe cognitive processes during a learning session that a (near) immediate post-test can suffice (Hulstijn, 2003). Wang, Thomas, and Ouellette (1992) pointed out that the results of a delayed post-test of recall can be contaminated by effects of retrieval practice in an immediate post-test unless immediate and delayed post-tests be taken by different experimental participants.

[7] The disparity, or heterogeneity, of the estimates of effect size across the four studies is relatively mild. For example, I^2 is only 2.8%, whereas it would be much closer to 100% if heterogeneity was likely to be a problem (Borenstein et al., 2009).

Acknowledgements

Two reviewers helped to make this a better article.

References

Abelson, R. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129–133. DOI: https://doi.org/10.1037/0033-2909.97.1.129

Asendorpf, J., Conner, M., De Fruyt, F., de Houwer, J., et al. (2013). Recommendations for increasing replicability in psychology. (Target article). *European Journal of Personality, 27*, 108–119. See also Authors' response, 138–144. DOI: https://doi.org/10.1002/per.1919

Baguley, T. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioural Sciences.* Basingstoke, UK: Palgrave Macmillan.

Bates, D., Mächler, M., Bolker, B., Walker, S., et al. (2017). R package lme4 Version 1.1–13: Linear Mixed-Effects Models using Eigen and S4. https://github.com/lme4/lme4/ and http://lme4.r-forge.r-project.org/.

Boers, F., & **Lindstromberg, S.** (2009). *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Basingstoke, UK: Palgrave-Macmillan. DOI: https://doi.org/10.1057/9780230245006

Boers, F., & **Lindstromberg, S.** (2012). Experimental and intervention studies of formulaic sequences in a second language. *Annual Review of Applied Linguistics*, *32*, 83–110. DOI: https://doi.org/10.1017/S0267190512000050

Boers, F., Lindstromberg, S., & **Eyckmans, J.** (2014). Is alliteration mnemonic without awareness-raising? *Language Awareness*, *23*(4), 291–303. DOI: https://doi.org/10.1080/09658416.2013.774008

Borenstein, M., Hedges, L., Higgins, J., & **Rothstein, H.** (2009). *Introduction to Meta-analysis*. Oxford: Wiley. DOI: https://doi.org/10.1002/9780470743386

Brandt, M., Ijerman, H., Dijksterhuis, A., Farach, F., Gellerd, J., et al. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. From: http://www.sciencedirect.com/science/article/pii/S0022103113001819.

Braver, S., Theommes, F., & **Rosenthal, R.** (2014). Continuously accumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342. DOI: https://doi.org/10.1177/1745691614529796

Breaugh, J. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29*, 79–97. DOI: https://doi.org/10.1177/014920630302900106

Cohen, J. (1994). The Earth is round, $p < 0.05$. *American Psychologist*, *49*, 997–1003. DOI: https://doi.org/10.1037/0003-066X.49.12.997

Cortina, J., & **Landis, R.** (2011). The earth is not round ($p = 0.00$). *Organizational Research Methods, 14*(2), 332–349. DOI: https://doi.org/10.1177/1094428110391542

Cumming, G. (2008). Replication and *p* intervals: *P* values predict the future only vaguely; but confidence intervals do much better. *Perspectives on Psychological*

Science, 3, 286–300. DOI: https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Hove: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29. DOI: https://doi.org/10.1177/0956797613504966

Dienes, Z. (2014). Using Bayes to make the most out of non-significant results. *Frontiers in Psychology.* DOI: https://doi.org/10.3389/fpsyg.2014.00781

Ellis, P. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-analysis, and the Interpretation of Research Results.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511761676

Eyckmans, J., & **Lindstromberg, S.** (2016). The power of sound in L2 vocabulary learning: Phonological similarity effects on the retention of conventionalized phrases. *Language Teaching Research, 21*(3), 341–361. DOI: https://doi.org/10.1177/1362168816655831

Field, A., Miles, J., & **Field, Z.** (2012). *Discovering Statistics using R*. Thousand Oaks, CA: Sage.

Forsberg, F. (2010). Using conventional sequences in L_2 French. *International Review of Applied Linguistics in Language Teaching, 48*, 25–51. DOI: https://doi.org/10.1515/iral.2010.002

Gass, S. (2009). A survey of SLA research. In: Ritchie, W., & Bhatia, T. (Eds.), *Handbook of Second Language Acquisition*, 3–28. Bingley, UK: Emerald.

Gries, S. (2011). Phonological similarity in multi-word symbolic units. *Cognitive Linguistics, 22*, 491–510. DOI: https://doi.org/10.1515/cogl.2011.019

Grissom, R., & **Kim, J.** (2012). *Effect Sizes for Research: Univariate and Multivariate Applications*, 2nd ed. Hove: Routledge.

Hulstijn, J. (2003). Incidental and intentional learning. In: Doughty, C. J., & Long, M. H. (Eds.), *The Handbook of Second Language Acquisition,* 349–381. Malden, MA: Blackwell Publishing.

Lai, J., Fidler, F., & **Cumming, G.** (2012). Researchers underestimate the variability of *p* values over replication. *Methodology, 8*, 51–62. DOI: https://doi.org/10.1027/1614-2241/a000037

Laufer, B., & **Waldman, T.** (2011). Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning, 61*, 647–672. DOI: https://doi.org/10.1111/j.1467-9922.2010.00621.x

Linck, J., & **Cunnings, J.** (2015). The utility and application of mixed-effects models in second language research. *Language Learning,* 65(Supplement 1), 185–207. DOI: https://doi.org/10.1111/lang.12117

Lindstromberg, S. Forthcoming. Surplus interword phonological similarity in English multiword units. *Cognitive Linguistics and Linguistic Theory.*

Lindstromberg, S. (2016). Inferential statistics in *Language Teaching Research*: A review and ways forward. *Language Teaching Research, 20*(6), 741–768. DOI: https://doi.org/10.1177/1362168816649979

Lindstromberg, S., & **Boers, F.** (2008). Phonemic repetition and the learning of lexical chunks: The power of assonance. *System*, *36*(3), 423–436. DOI: https://doi.org/10.1016/j.system.2008.01.002

Lindstromberg, S., & **Eyckmans, J.** (2014). When does assonance make lexical phrases memorable? *European Journal of Applied Linguistics and TEFL*, *3*(1), 93–107.

Long, M. (1983). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition, 5*, 177–193. DOI: https://doi.org/10.1017/S0272263100004848

Mackey, A. (2013). Methodology in SLA research: Past, present and future. Plenary presentation. The 23rd conference of the European Second Language Association, 28–31. August. University of Amsterdam.

McClelland, G. (1997). Optimum design in psychological research. *Psychological Methods, 2*, 3–9. DOI: https://doi.org/10.1037/1082-989X.2.1.3

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166. DOI: https://doi.org/10.1037/0033-2909.105.1.156

Nekrasova, T. (2009). English L_1 and L_2 speakers' knowledge of lexical bundles. *Language Learning, 59*, 647–686. DOI: https://doi.org/10.1111/j.1467-9922.2009.00520.x

Norris, J. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 56*, 97–126. DOI: https://doi.org/10.1111/lang.12114

Nosek, B., Aarts, A., Anderson, C., Anderson, J., et al. (2015). Estimating the reproducibility of psychological science. (Open Science Collaboration). *Science*, *349* (6251). DOI: https://doi.org/10.1126/science.aac4716

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35*, 655–87. DOI: https://doi.org/10.1017/S0272263113000399

Plonsky, L., & **Gass, S.** (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*, 325–66. DOI: https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & **Oswald, F.** (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. DOI: https://doi.org/10.1111/lang.12079

Polio, C. (2012a). Replication in published applied linguistics research: A historical perspective. In: Porte, G. (ed.), *Replication Research in Applied Linguistics*, 47–91. Cambridge: Cambridge University Press.

Polio, C. (Ed.) (2012b). *Annual Review of Applied Linguistics*, *32*.

Porte, G. (2012a). Introduction. In: Porte, G. (Ed.), *Replication Research in Applied Linguistics*, 1–17. Cambridge: Cambridge University Press.

Porte, G. (Ed.) (2012b). *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Clarendon Press.

Tversky, A., & **Kahneman, D.** (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110. DOI: https://doi.org/10.1037/h0031322

Valentine, J., Pigott, T., & **Rothstein, H.** (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*, 215–247. DOI: https://doi.org/10.3102/1076998609346961

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. DOI: https://doi.org/10.18637/jss.v036.i03

Wang, A., Thomas, M., & **Ouellette, J.** (1992). Keyword mnemonic and retention of second-language vocabulary words. *Journal of Educational Psychology, 84*, 520–528. DOI: https://doi.org/10.1037/0022-0663.84.4.520

Westfall, J., Kenny, D., & **Judd, C.** (2014). Statistical power and optimum design in experiments in which participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. DOI: https://doi.org/10.1037/xge0000014

Wilcox, R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed. London: Academic Press.

Wilcox, R. (2017). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*, 2nd ed. Boca Raton, FL: CRC Press.

How to cite this article: Lindstromberg, S. and Eyckmans, J. (2017). The particular need for replication in the quantitative study of SLA: A case study of the mnemonic effect of assonance in collocations. *Journal of the European Second Language Association*, 1(1), 126–136, DOI: https://doi.org/10.22599/jesla.26

Submitted: 22 January 2017 **Accepted:** 30 June 2017 **Published:** 01 August 2017

JESLA Journal of the European Second Language Association

Journal of the European Second Language Association is a peer-reviewed open access journal published by White Rose University Press.

OPEN ACCESS